IN THE UNITED STATES DISTRICT COURT
FOR THE SOUTHERN DISTRICT OF TEXAS
CORPUS CHRISTI DIVISION

MARC VEASEY, *et al.*, Plaintiffs, v. RICK PERRY, *et al.*, Defendants.	Civil Action No. 2:13-cv-193 (NGR)
UNITED STATES OF AMERICA, Plaintiff, TEXAS LEAGUE OF YOUNG VOTERS EDUCATION FUND, *et al.*, Plaintiff-Intervenors, TEXAS ASSOCIATION OF HISPANIC COUNTY JUDGES AND COUNTY COMMISSIONERS, *et al.*, Plaintiff-Intervenors, v. STATE OF TEXAS, *et al.*, Defendants.	Civil Action No. 2:13-cv-263 (NGR)

TEXAS STATE CONFERENCE OF NAACP BRANCHES, *et al.*, Plaintiffs, v. NANDITA BERRY, *et al.*, Defendants.	Civil Action No. 2:13-cv-291 (NGR)
BELINDA ORTIZ, *et al.*, Plaintiffs, v. STATE OF TEXAS, *et al.*, Defendants	Civil Action No. 2:13-cv-348 (NGR)

DECLARATION OF YAIR GHITZA

Pursuant to 28 U.S.C. § 1746, I, Yair Ghitza, make the following declaration:

Report on Statistical Estimation of the Race of Individual Registered Voters in Texas

1. My name is Yair Ghitza. I am the Chief Scientist at Catalist, LLC, a data services company that collects, standardizes, and enhances data from official voter registration databases, as well as other commercial, public, and private data sources. I have over 10 years of experience in statistics, political science, and computer science in both professional and academic settings. At Catalist, I am mainly responsible for constructing and overseeing the construction of thousands of predictive models. Usually, these models predict some characteristic, attitude, or likely behavior of the people in the database. I have built these types of models for a wide range of topics, from likelihood of voting in a particular election, to likelihood of having children in the household, conditional on a wide range of other data points. I also lead the research efforts of Catalist, often building new statistical methods to deal with different types of data, and developing new methods of leveraging the data to help our clients achieve their goals.

2. I successfully defended my doctoral dissertation in March 2014 and will shortly be receiving a PhD from Columbia University in Political Science. My dissertation—*Applying Large-Scale Data and Modern Statistical Methods to Classical Problems in American Politics*—was accepted with distinction and is currently nominated for the annual Savage Award in Applied Methodology, awarded to a dissertation that makes outstanding contributions in the field of applied Bayesian statistics. My main areas of expertise are statistical methods and American politics, particularly focusing on estimating public opinion. Prior to my time at Columbia, I was a visiting research assistant in the Media Lab at MIT, performing core research in artificial intelligence and computer vision.

3. My academic work has appeared in journals such as the American Journal of Political Science; The Forum; Statistics, Politics, and Policy; and Proceedings of the IEEE International Symposium on Wearable Computers. It has also been featured at invited conferences such as the Annual State Politics and Policy Conference; the Annual Meeting of the Society for Political Methodology; the Annual Meeting of the American Political Science Association; the Annual Conference of the Public Choice Society; the Annual Conference of the European Political Science Association; and the Mapping Science Committee at the National Academy of Sciences. My C.V. is attached to this report.

4. Catalist was compensated at a rate of $5.00 per thousand voters to provide the data regarding Texas registered voters discussed in this report—specifically, racial identification estimates along with geographic and other data, such as whether this voter is deceased, as discussed in this report. For my work in preparing this report and testifying at trial, Catalist is compensated at a rate of $150 per hour. I have not previously provided expert testimony in any litigation.

5. Catalist is a data utility that provides services to civic engagement and advocacy organizations as well as political campaigns. Catalist compiles, enhances, stores, and updates person-level data for the entire U.S. adult population. Additionally, Catalist provides tools, services, analyses, and expertise to facilitate planning, analyzing, and executing data-driven voter contact and other civic advocacy programs.

6. Catalist maintains a national database of voting-age persons, the data for which is obtained from official voting rolls in all 50 states and the District of Columbia, as well as from national commercial consumer databases. Combining these datasets with publicly-available data, such as Census data, and private data from our clients and business partners, results in a national database that contains nearly a thousand attributes, giving Catalist a rich, national database of civic and commercial behavior that is updated state-by-state on an ongoing and periodic basis, with multiple updates of state data occurring as frequently as every week. These updates may include new information on individuals' voter registration status, voting history, official political and administrative districts, deceased records, household income level, residence changes, and hundreds of other attributes, some of which are estimated through statistical models. Catalist then augments its database with a number of modeled predictions about each person's likely civic behavior or preferences on a number of subjects relevant to civic participation.

7. For several states such as South Carolina, Louisiana, Florida and North Carolina, data from official voter rolls includes information about a voter's self-reported race, which is collected when the voter registers to vote for the first time or updates his/her registration. The Catalist database includes this race information from states that collect it from voters. When self-reported race is unavailable, Catalist uses statistical estimates based on other attributes to estimate voters' race.

8. Catalist purchases these estimates from vendors who are experts in creating statistical models specifically to predict voter race and ethnicity. In 2009, Catalist conducted an extensive comparison of the products offered by several commercial race coding vendors. Catalist measured the accuracy and coverage of the various vendors by comparing them against a large sample of self-reported race from voter rolls, surveys and other person-level contact programs that Catalist has collected since 2006.

9. Based on this initial analysis, Catalist selected the data vendor CPM Technologies as the source of Catalist's modeled race data. CPM race coding has since been re-validated regularly by Catalist to measure continued race modeling accuracy, again through the use of self-reported race from both officially sourced voter files and data from hundreds of polls conducted by Catalist's clients that asked survey respondents a racial demographic question. This allows Catalist to compare CPM's modeled race prediction for an individual with the individual's own, self-reported racial identity. The most recent re-validation is described in paragraph 15.

10. In Catalist's database, where self-reported race from official voting rolls is available, that self-reported race is used. Where race information is not available, Catalist uses CPM Technologies race coding algorithms to predict race.

11. CPM's race coding algorithm uses a multi-layered, tree based approach. This means that the algorithm looks at combinations of demographic and other information (e.g., first name, location, sex) and assigns the most likely race based on those combinations. To find out which were the most likely combinations, CPM "trained" the model using data where all the information including race were known. CPM "retrains" the statistical model once a year on average with updated data. Each time CPM retrains the model, it provides Catalist with a new program to assign the predicted race.

12. Every year, CPM trains the model using census data, voter file data where permissible to be used for commercial purposes, and data from various commercial vendors. CPM does external validation on the models using self-reported survey responses and self-reported race records from official voter databases that does not overlap with the training set of voter file race data, ensuring that the models generalize to the full population instead of just the data that was used to build the model. Both tree-

based modeling and this method of validation are in line with widely accepted standards and practices in predictive modeling and analytics.

13. The CPM algorithm assigns a race to each record along with a "race confidence" value, a measure of the accuracy CPM ascribes to a given modeled race value. These race confidence values are segmented by CPM Technologies in to the following values, ordered from most confident in the appended race to least confident:

    1. Highly Likely
    2. Likely
    3. Possibly

    Catalist records may also carry no race confidence value if the race value is sourced from the voter file or records may be labeled with a race confidence value of "Uncoded" if race value was neither available from official voting rolls or the CPM Technologies algorithm was unable to assign any single race value to the record with sufficient confidence.

14. Catalist has compared the predicted race from the algorithm with the self-reported race from voter rolls where that data is available. For records with the highest race confidence scores, Catalist has found that the predictions match the voters' self-reported race with 90% accuracy or greater in most cases. This relationship between confidence scores and accuracy is covered more fully in paragraph 15.

15. Catalist most recently validated CPM race coding results against the self-reported race available on certain voter files in early 2014. A topline of that validation is shown in the table below. Here, we applied CPM's race coding technology to database records in nine states that include self-reported race on the official voter rolls: Alabama, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, and Tennessee. We compared CPM's predictions (as seen in each row of the table) to self-reported race, among people who did self-report their race, in order to determine how accurate CPM's predictions were.

		CPM Race Confidence Percentage				CPM Percent Correct		
Race	N_size	Highly Likely	Likely	Possibly	Uncoded	Highly Likely	Likely	Possibly
White	21315485	70%	21%	7%	2%	97%	91%	78%
Black	8029434	40%	19%	42%	0%	93%	81%	54%
Hispanic	2206470	47%	21%	32%	0%	90%	77%	58%
Asian	282925	41%	21%	38%	0%	90%	76%	39%
Native American	74560	67%	10%	22%	0%	15%	69%	30%
Other	110989	0%	0%	0%	100%	NA	NA	NA
Unknown	32336	0%	0%	0%	100%	NA	NA	NA

16. The race data on the Catalist file has been used and relied on in academia regularly. Recent academic studies that used Catalist's race data include the following:

    - Mann, Christopher B., and Casey A. Klofstad. "The Role of Call Quality in Voter Mobilization: Implications for Electoral Outcomes and Experimental Design." *Political Behavior* (2010).
    - Ansolabehere, Stephen, and Eitan Hersh. "Validation: What Big Data Reveal About Survey Misreporting And The Real Electorate." *Political Analysis* (2012).
    - Ansolabehere, Stephen, and Eitan Hersh. "Gender, Race, Age and Voting: A Research Note." *Politics and Governance* (2013).

- Enos, Ryan D., and Anthony Fowler. "The Effects of Large-Scale Campaigns on Voter Turnout: Evidence from 400 Million Voter Contacts." Working Paper (2013).
- Fraga, Bernard L. "Winning the Race, Losing the Base? Demobilization, Competitiveness, and Electoral Influence." Working Paper (2013).
- Hersh, Eitan and Clayton Nall. "A Direct-Observation Approach to Identify Small-Area Variation in Political Behavior: The Case of Income, Partisanship, and Geography." Working Paper (2013).
- Hersh, Eitan D., and Brian F. Schaffner. "Targeted Campaign Appeals And The Value Of Ambiguity." *The Journal of Politics* (2013).
- Rogers, Todd, and Masahiko Aida. "Vote Self-prediction Hardly Predicts Who Will Vote, And Is (Misleadingly) Unbiased." *American Politics Research* (2014).
- Fraga, Bernard L. "Assessing the Causal Impact of Race-Based Districting on Voter Turnout." Working Paper (2014).
- Fraga, Bernard L. "Candidates or Districts? Reevaluating the Role of Race in Voter Turnout." Working Paper (2014).
- Ashok, Vivekinan, et al. "Dynamic Voting in a Dynamic Campaign: Three Models of Early Voting." Working Paper (2014).

Some findings from these academic papers that were specific to Catalist race coding include the Hersh and Nall (2013) paper, which used Catalist race data to look at differences in regional voting based on income. They found the Catalist race data to be an important part of the trends discovered in this research. Ansolabehere and Hersh (2012) included a validation of Catalist race data against 2008 online survey panels that yielded a strong alignment. Hersh and Schaffner (2013) also report validation statistics that are in the expected range. As some of these academic papers have discussed, Catalist has a vested interest in assuring that the race predictions it maintains in its database and provides to clients are as accurate and unbiased as possible.

17. Catalist's database also indicates whether persons on official voter rolls are believed to be deceased. Catalist obtains death information from multiple sources, both official and commercial. Catalist subscribes to the Social Security death master file, which is updated monthly, and those updates are matched and merged into the Catalist deceased indicator. Catalist also sources deceased information from Interactive Marketing Solutions, a company that maintains a Recently Recorded Deceased File and a Deceased Do Not Call list. These two files are updated monthly, and those updates are also matched and merged in to the Catalist deceased indicator.

18. Furthermore, Catalist's database indicates whether persons on a state's official voter rolls have submitted a National Change of Address (NCOA) to the United States Postal Service. An NCOA submission is an individual indicating to the USPS that they want mail sent to them at an address to be forwarded on to a new address. Catalist runs NCOA on its national file on a rolling, state-by-state schedule. This schedule is set in such a way that no state goes more than 90 days without being submitted for NCOA, processed, and re-released with newly updated NCOA data. Due to restrictions in the storage of NCOA data, these official NCOA indicators are dropped off of the database after 18 months.

19. Catalist's database also indicates whether persons on a state's official voter rolls are modeled to be "deadwood." The deadwood indicator identifies records that are probably deceased, no longer at the recorded address or otherwise not accurately recorded. The deadwood model uses the deceased and NCOA indicators described in the previous sections, in conjunction with other factors such as vote history and voter status collected from official voting rolls, to assign a deadwood category to all voter records on the Catalist file. This model is applied every time Catalist processes a voter file from a secretary of state. The deadwood model can be considered a modeled indicator, estimating the

likelihood that the person listed in the record is not a living, eligible voter, who resides in the listed address. The deadwood model is not purely cumulative of the deceased and NCOA flags.

20. Counsel for the Department of Justice provided Catalist with a database of Texas registered voters. There were 13,564,420 voter records in the file DOJ provided to Catalist.

21. Upon receipt of this file from the Department of Justice, Catalist took the following steps to match those records to our database of Texas registered voters: a) standardized the file format using address standardization and CASS correction;[1] b) removed malformed records;[2] c) applied NCOA processes as described above; d) geocoded the addresses for mapping to census geography; e) ran gender imputation logic where gender was missing on file; f) ran CPM race coding because race is not present in the official list of Texas registered voters maintained by the Texas Secretary of State; g) appended deceased flags as described above; h) appended official district and jurisdiction data; i) scored and appended the deadwood model described above.

22. Catalist provided the Department of Justice with information on those approximately 13.5 million Texas registered voters. This dataset included the race coding data and whether any of those voters are considered likely to be deceased, deadwood, or have a change of address NCOA flag.

23. 50.20% of the Texas voter records had a confidence level of "highly likely" for the race estimate, 22.52% were "likely," 25.08% were "possibly" and 2.20% were uncoded.

24. 323,620 out of the approximately 13.5 million Texas voter records returned were marked by Catalist as deceased.

25. Out of the approximately 13.5 million Texas voter records returned, 236,429 were marked by Catalist as possible deadwood, 34,435 were marked as probable deadwood, and approximately 13.3 million were marked as not deadwood.

26. 882,113 out of the approximately 13.5 million Texas voter records returned were marked by Catalist with an NCOA flag.

27. The numbers reported in paragraphs 23 through 26 are not unusual for states in the Catalist database.

[1] CASS is an address certification system offered by the U.S. Postal Service that improves the accuracy of carrier route, 5-digit ZIP, ZIP + 4, and delivery point codes that appear on mail pieces.
[2] For example, records that were missing first or last name.

I declare under penalty of perjury that the foregoing is true and correct.
Executed this 27th day of June, 2014.

/s/
Yair Ghitza

Yair Ghitza

365 N. Halsted St., Apt 801
Chicago, IL 60661

202-538-0670
yghitza@gmail.com

EDUCATION

Columbia University — **New York, NY**
Ph.D. Forthcoming, Department of Political Science — *September 2007 – May 2014 (Defended, not yet Deposited)*
M.S. Political Science, May 2009
M.Phil. Political Science, December 2010

- Focus on the study of American elections, specializing in the use of quantitative methods, including Statistics, Data Mining, Machine Learning, and Data Visualization.
- Developed statistical procedure for estimating small population subgroup estimates from national polls (Multilevel Regression and Poststratification).
- Conducted various (published and non-published) other analyses on public opinion and political behavior, leveraging survey, demographic, geographic, and large scale voter registration data.
- Teaching experience in introductory statistics for graduate students.

University of Michigan — **Ann Arbor, MI**
BSE, Computer Science and Engineering, Magna Cum Laude — *September 1999 – April 2003*

SELECTED WRITING

- Ghitza, Yair, and Andrew Gelman. 2013. "Deep Interactions with MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups." *American Journal of Political Science* 57(3):762-776.
- Erikson, Robert and Yair Ghitza. 2012. "Setting the Agenda Setter." Prepared for 2012 Meeting of the American Political Science Association, New Orleans, LA. August 30-September 2, 2012.
- Ghitza, Yair. "Who's Going to Vote? Demographic Turnout Forecasting Using Pre-Election Polls." Working Paper.
- Erikson, Robert, Yair Ghitza, and Christopher Wlezien. 2010. "Differential Campaign Effects in Battleground and Non-Battleground States? An Analysis of Recent Presidential Elections." Presented at the Annual State Politics and Policy Conference, Springfield, Illinois, June 3-5, 2010.
- Gelman, Andrew, Daniel Lee, and Yair Ghitza. 2010. "Public Opinion on Health Care Reform." *The Forum* 8(1).
- Gelman, Andrew, Daniel Lee, and Yair Ghitza. 2010. "A Snapshot of the 2008 Election." *Statistics, Politics, and Policy* 1(1).
- Gelman, Andrew, Jonathan P. Kastellec, and Yair Ghitza. 2009. "Beautiful Political Data." In Beautiful Data: The Stories Behind Elegant Data Solutions. O'Reilly Media.
- Ghitza, Yair, and Todd Rogers. 2009. "Data Driven Politics". In The Change We Need: What Britain Can Learn from Obama's Victory. Ed. Nick Anstead and Will Straw. Fabian Society.
- Roy, Deb, Yair Ghitza, Jeff Bartelma, and Charlie Kehoe. 2004. "Visual Memory Augmentation: Using Eye Gaze as an Attention Filter." Proceedings of the IEEE International Symposium on Wearable Computers.

SELECTED WORK HISTORY

Catalist / Copernicus Analytics — **Washington, DC**
Senior Scientist / Consultant — *August 2004 – Present*

- Created and implemented large-scale statistical models and data visualization projects for national and dozens of state and Congressional campaigns.
- This "microtargeting" work predicted probability of candidate support, voter turnout, issue support, and financial donation on the individual level for hundreds of millions of registered voters.

MIT Media Lab, Cognitive Machines Group — **Cambridge, MA**
Visiting Research Assistant — *July 2003 – May 2004*

- Performed core research in artificial intelligence and computer vision.
- Designed "attention glasses," a wearable attention aid that augments visual search capability.
- Currently developing an interface that improves communication abilities of paralyzed patients.